**Learning Journal Two**

**S. Hayes**

**June 8**

***How can we characterize the center and distribution of data values?* Chapters 4, 5, 6**

**Mean versus median.** The *mean* is another way of determining the center of a distribution of values. It is calculated by adding up all the values and dividing by the total number of values. The mean is best used when a distribution is symmetrical. This compares to the median, which only considers the order of the values to find the central point. The *median* is best used in cases where a distribution is skewed because it is resistant to values that are extraordinarily large or small because it ignores their distance from the center point.

***Dispersion****.* This statistic shows the amount of *variation or spread* in the values of a variable. There are three ways to measure spread.

*Range.* One approach is to examine the *range* of the data by subtracting the maximum value from the minimum value.

*IQR.* A better approach is to identify the *Interquartile range* (IQR) the area in the distribution which accounts for half the data located above the lower (25th) and below the upper (75th) quartiles. To calculate IQR, subtract the lower quartile from the upper quartile (Q3-Q1=IQR). These are best used when a distribution is skewed by outliers. Note also that the mean is typically paired with standard deviation and median with IQR.

*Standard Deviation.* Unlike the IQR which only examines the middle range of values, a third approach to measuring spread is to calculate the standard deviation, which provides details on how far away *each* value is from the mean. The standard deviation is best used for symmetrical distributions. Because it can be used to measure distances from the mean, the standard deviation also provides a way to standardize or compare performances based on different scales. A z-score can be calculated by subtracting the mean from the value of interest and dividing it by the standard deviation. For example, a z-score of 2 tells us that a data value is two standard deviations above the mean, and allows us to use this as a comparison again another measure that has also been standardized as a z-score.

***What is the relationship between standard deviation and the Normal model?***

**Normal Model.** The standard Normal model, in which the mean is 0 and a standard deviation of 1, is best used to describe unimodal and symmetric distributions whose shape matches a bell-shaped curve. In this model 68% of all values fall within 1 standard deviation of the mean, 95% fall within 2 standard deviations, and 99.7% of all values fall within 3 standard deviations. Using these guidelines it is possible to compare relative performance of quantitative variables using z-scores and normal percentiles.

**June 10**

**Scatter plots, associations and correlation (Chapter 7)**

So far we have been looking at analyses involving just one quantitative variable. Correlation measures the strength of *linear* association between two quantitative variables. Scatter plots are used to depict the form (straight, curved, circle), direction (positive, negative, neutral) and strength (how much scatter) of the relationship between two variables. The x variable, on the horizontal axis, represents the explanatory variable. The y variable, on the vertical axis, represents the response variable.

The correlation coefficient “r” can range in value from -1.0 to +1.0. A high positive correlation between two variables means that when one is high, so is the other. A perfect negative correlation means that when variable A is high, B is low, and vice versa. A zero correlation means there is no relationship between the two variables.

Correlation coefficients are very sensitive to outliers and for this reason we should always run a scatter plot to determine if the data is linear. The authors also caution that using correlation tables instead of scatter plots to make determinations of correlation may be misleading because they provides no visual check for linearity and outliers. In cases where the scatter plot is slightly curved, it may make sense to re-express the variables by to see if this straightens the line. Kinnear provides a good explanation of this on p. 121.

Both Kinnear and Vella et al. also warn that correlation is not to be confused with causation. At best we can say is that there is a “relationship” or “association” between the two variables. It is possible that a lurking or hidden variable is simultaneously affecting both variables, and may be the actual cause of the relationship.

Questions

I keep returning to the formulas for calculating standard deviation and correlation. There are many common elements: calculating the deviation from the mean for the y variable (y – y bar) and the x variable (x- xbar); dividing by the total count of values (n-1), and; the use of the calculated standard deviation for variables, sx and sy. Would it be meaningful to take the values of all the x’s and all the y’s in a scatter plot and calculate the standard deviation (assuming the distribution matches the Normal model) for both sets separately and then overlay both on a normal curve? What would I see?

**June 11**

**Sample Surveys (Chapter 12)**

In quantitative research, a *population* is the group to which a researcher wishes to generalize his or her results. This may be a small group of people or a very large one measuring millions. In contrast, a sample is defined as a subset of a population. The use of a random sample ensures that, on average, the sample has the same characteristics of the population under study. If the statistics that are computed from a sample of a larger population accurately reflect its corresponding parameters, the sample is said to be *representative* of the population.

A simple random sample (SRS) is used to ensure that each combination of people that comprise the population have an equal chance of being selected. The sampling frame is a list from which the sample is drawn. Sampling variability accounts for different people being selected each time a random sample is drawn. There are many ways to draw samples from a population. Stratified random sampling divides a population into homogenous groups and uses random sampling to identify participants from each group. This allows for matching proportions of characteristics within a larger population. Cluster sampling is used when heterogeneous slices of the population are required. This method is used when it is necessary to be cost effective or practical. Multistage sampling combines several methods. Systematic sampling takes place when every nth unit is sampled. Typically the starting point is chosen at random.

There is no way to correct for badly considered sampling decisions that introduce bias into the sample design. Relying on volunteers, settling for easily accessible convenience samples, selecting inappropriate sample frames, and overlooking population segments can all undermine sample quality.

SPSS

I attempted to recode all of the *Refused to Answer* responses to 99. When I did this, I was left with many cases where I needed to go back and manually ‘fix’ other codes that were affected by this change. I tried this twice and in both cases the same thing happened. What am I doing wrong? Is there a better way to do this?

DeVeaux, R., Velleman, P., & Block, D. (2009). Intro Stats, 3rd ed., Boston: Pearson Education.

Kinnear, P.R.& Gray, C. D. (2010*). IBM SPSS Statistics 18 Made Simple*. New York: Psychology Press.